Group 7

Body Fat Prediction

Contents

Introduction

Background and Data Cleaning

Data cleaning

- **Scale**
  We changed the units of weight and height to kg and cm because all of the other variables in the dataset are measured in cm.

- **Height**
  We fixed a person who is only 74.94 cm tall using the relationship between bmi, weight and height (bmi=kg/m^2).

- **Bodyfat**
  We removed the person with 0% body fat percentage, because this is scientifically impossible and we couldn't manage to fix this.

Model Selection and Results

Model Selection, Visualization, Result Discussion

Models Selection

- **Stepwise linear regression model without interaction**
  Weight, age,neck,abdomen,thigh,forearm,wrist are significant
- **Lasso regression**
  Age,height,neck,abdomen,forearm,wrist are significant

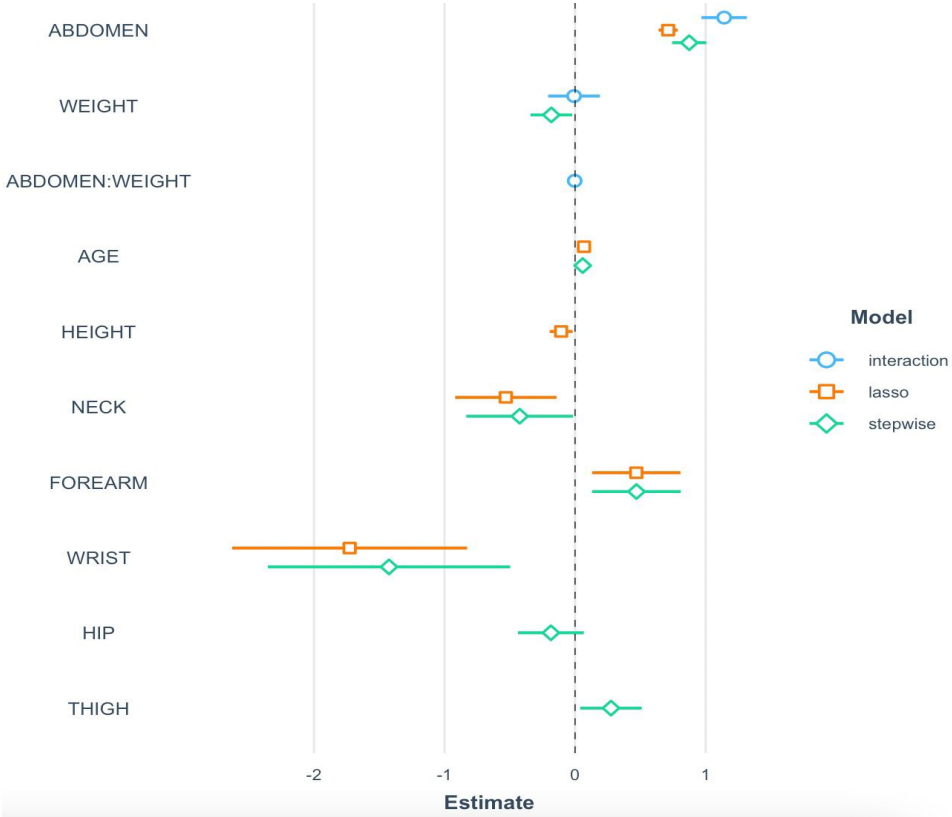
There are too many predictors and we want a simpler model

1. The two models above show that weight, age, abdomen, height are four predictors significant and easy to measure
2. We fit a linear model only consider these four predictors, and find that only abdomen and weight are significant
3. We suggest that there are interaction effect between abdomen and weight
4. So we fit a linear model with interaction

- **Linear regression with interaction**
  Abdomen and the interaction between abdomen and weight are significant

Visualization of Results

- This plot shows the 95% CI of all the significant predictors in our three models
- We can see that estimates for each predictors are close to each other



Discussions of results

- MSE and R^2 are similar in these three models
- Overall p-value are all smaller than 0.05
- Among them, the interaction model uses least amount of predictors and is the simplest one

Model	MSE	R^2	Overall p-value
stepwise	15.12	0.74	<2e-16
Interaction model	16.08	0.72	<2e-16
lasso	15.67	0.73	<2.2e-16

Final Model Interpretation

Final Model, Visualization, Model Diagnostics
, Strengths and Weaknesses

Final model

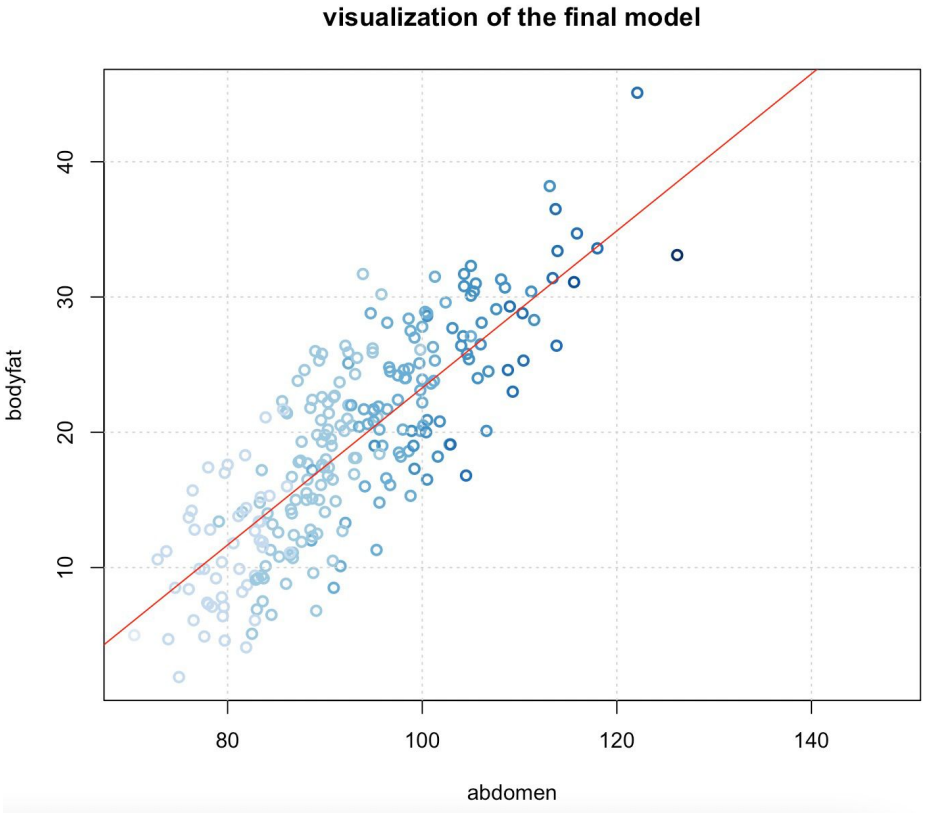
bodyfat=-64.4-0.008weight+1.14abdomen-0.003weight*abdomen

- **Michael Phelps Example: weight(87 kg), abdomen(86.36 cm)**
  Michael's estimated body fat percentage is 8.33%
  This result is pretty close to what we got from Google, which is 10.3%.

- **For A man who has 89 cm abdomen circumference**
  our model predicts that his body fat percentage will decrease 0.275% for every 1 kg increase in his weight.

- **For a man who has 70 kg weight**
  Our model predicts that his body fat percentage will increase 0.93% for every 1 cm increase in his abdomen circumference.

Visual description of final model



- We draw a scatterplot of abdomen and bodyfat
- Red line represents the fitted line of bodyfat and abdomen. When abdomen increases, bodyfat increases
- Color of the points represents the value of interaction of abdomen and weight
- Color the deepest in lower right side

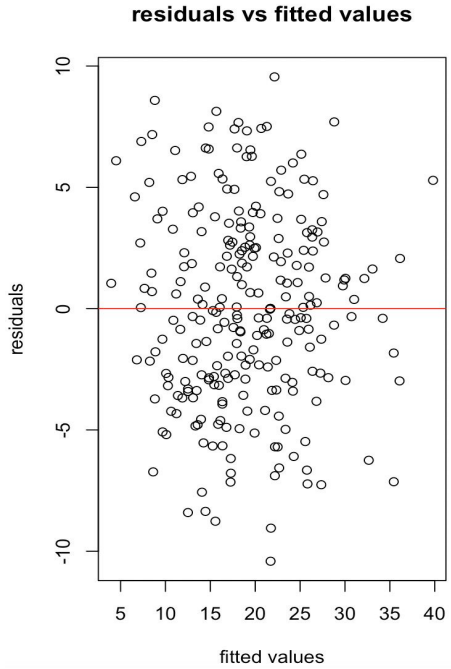
Statistical properties of final model

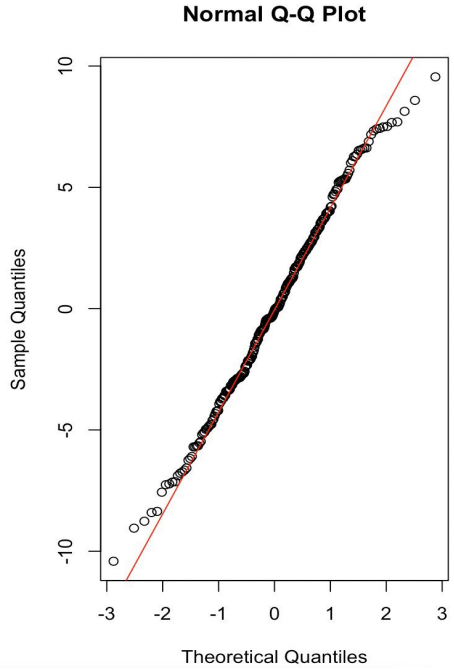
predictor	coefficient	p-value	95% lower CI	95% upper CI
intercept	-64.4	4.63e-15	-79.60	-49.25
weight	-0.008	0.94	-0.21	0.19
abdomen	1.14	<2e-16	0.97	1.32
weight*abdomen	-0.003	0.001	-0.005	-0.001

- Abdomen and the interaction term weight*abdomen are significant under alpha=0.05
- Overall model is significant with p-value<2e-16
- R^2=0.72, MSE=16.08

Model diagnostics

- We test the normality and randomness of the residuals
- Q-Q plot shows that residuals are approximately normally distributed
- Residuals vs fitted values plot shows that residuals have mean 0 and show no obvious pattern





Strengths and weaknesses

bodyfat=-64.4-0.008weight+1.14abdomen-0.003weight*abdomen

- **Strengths**

  Simple and Easy to explain

  Residuals normally distributed

- **Weaknesses**

  Only two predictors

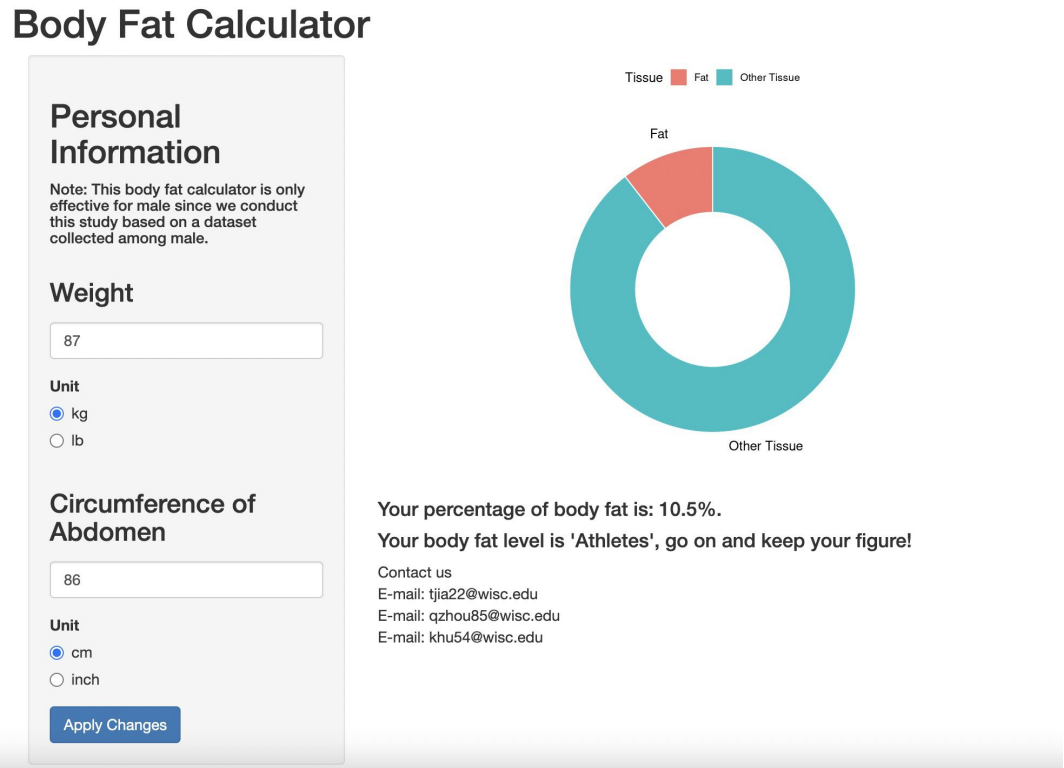
  Data is inaccurate for weight and abdomen

  Outliers in normal Q-Q plot

Shiny App

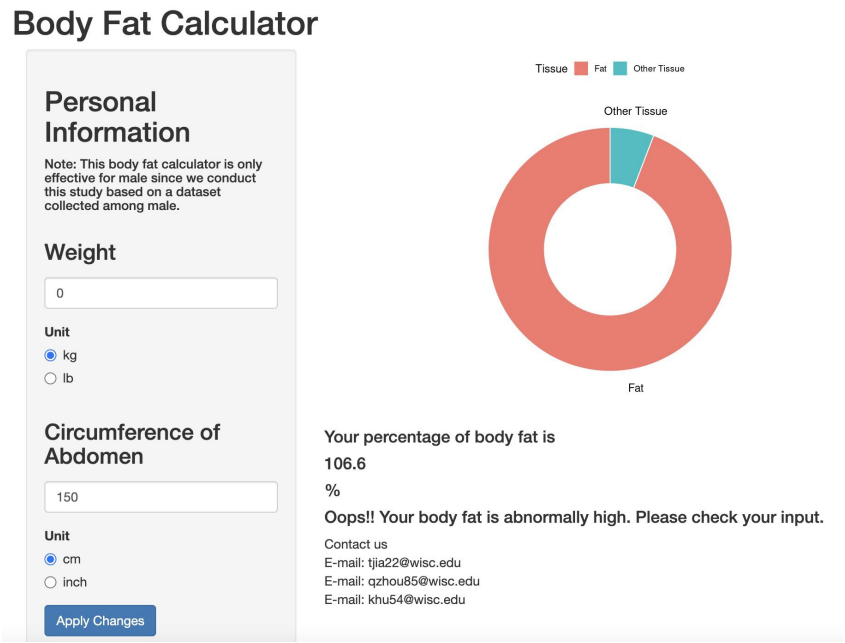
Michael Phelps Example, Extreme Values

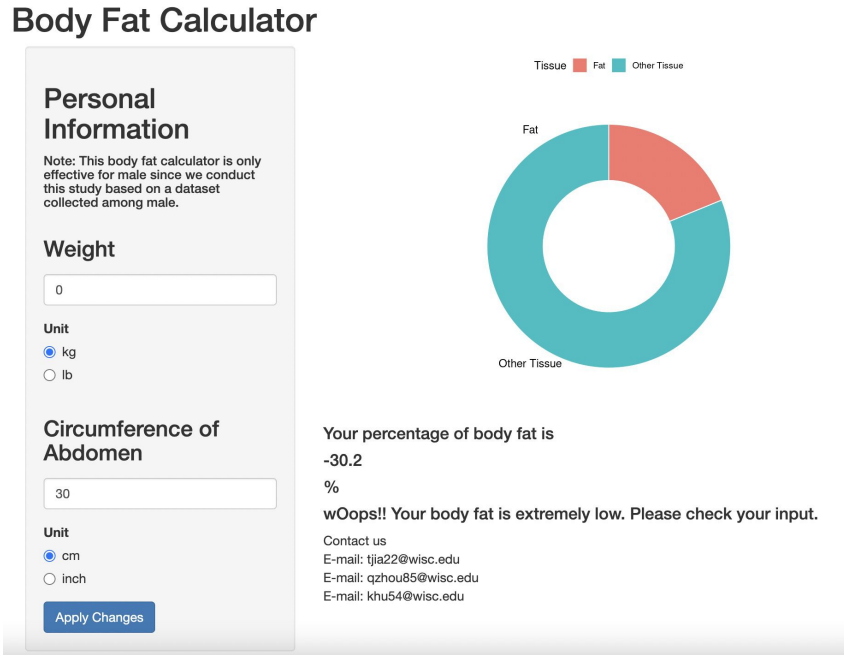
Michael Phelps Example



- We use Michael Phelps's data to introduce our shiny app
- The outcome is 10.5%, which is really close to the Google result 10.3%
- There is some difference between our predicted data and the shiny app outcome due to rounding

Extreme Values





Thank you

Contribution: